ABSTRACT OF THE DISCLOSURE

Improved duplicate and near-duplicate detection techniques may assign a number of fingerprints to a given document by (i) extracting parts from the document, (ii) assigning the extracted parts to one or more of a predetermined number of lists, and (iii) generating a fingerprint from each of the populated lists. Two documents may be considered to be near-duplicates if any one of their fingerprints match.